小規模誤りデータからの日本語学習者作文の助詞誤り訂正

今村　賢治[†]・齋藤　邦子[†]・貞光　九月[†]・西川　　仁[†]

本稿では，置換，挿入，削除操作を行う識別的系列変換で日本語学習者作文の助詞誤りを自動訂正する．誤り訂正タスクの場合，難しいのは大規模な学習者作文コーパスを集めることである．この問題を，識別学習の枠組み上で 2 つの方法を用いて解決を図る．一つは日本語としての正しさを測るため，少量の学習者作文から獲得した n-gram 二値素性と，大規模コーパスから獲得した言語モデル確率を併用する．もう一つは学習者作文コーパスへの直接的補強として，自動生成した疑似誤り文を訓練コーパスに追加する．さらに疑似誤り文をソースドメイン，実際の学習者作文をターゲットドメインとしたドメイン適応を行う．実験では，n-gram 二値素性と言語モデル確率を併用することで再現率の向上ができ，疑似誤り文をドメイン適応することにより安定した精度向上ができた．

キーワード：助詞誤り訂正，言語モデル，疑似誤り生成，ドメイン適応，識別的系列変換

Particle Error Correction of Japanese Learners from Small Error Data

Kenji Imamura[†], Kuniko Saito[†], Kugatsu Sadamitsu[†] and Hitoshi Nishikawa[†]

This paper presents grammatical error correction of Japanese particles written by foreign Japanese learners. Our method is based on discriminative sequence conversion, which corrects particle errors by substitution, insertion, or deletion. For this kind of error correction task, it is difficult to collect large learners' corpora. We attempt to solve this problem based on a discriminative learning framework which uses the following two methods. First, language model probabilities obtained from large Japanese corpora are combined with n-gram binary features obtained from the learners' corpora. This method is applied in order to measure the correctness of Japanese sentences. Second, automatically generated pseudo-error sentences are added to the learners' corpora in order to enrich the corpora directly. Furthermore, we apply domain adaptation, in which the pseudo-error sentences (the source domain) are adapted to the real-error sentences (the target domain). Experimental results show that the recall rate has been improved by using both the language model probabilities and the n-gram binary features. Stable improvement has been achieved by using pseudo-error sentences with the domain adaptation.

Key Words: *Particle Error Correction, Language Models, Pseudo-Error Generation, Domain Adaptation, Discriminative Sequence Conversion*

[†] 日本電信電話株式会社，NTT メディアインテリジェンス研究所, NTT Media Intelligence Laboratories, NTT Corporation

1 はじめに

日本語学習者の作文の誤り訂正は，教育の一環としてだけでなく，近年はビジネス上の必要性も生じてきている．たとえば，オフショア開発（システム開発の外国への外部発注）では，中国，インドなどへの発注が増加している．外国に発注する場合，日本との意思疎通は英語または日本語で行われるが，日本語学習者の多い中国北部では，日本語が使われることも多い．しかし，中国語を母語とするものにとって日本語は外国語であり，メールなどの作文には誤りを含み，意思疎通に問題となるため，それらを自動検出・訂正する技術が望まれている (大木他 2011; 末永，松嶋 2012)．そこで本稿では，日本語学習者作文の誤り自動訂正法を提案する．外国人にとって，助詞はもっとも誤りやすい語であるため，本稿では助詞の用法を訂正対象とする．

日本語の助詞誤り訂正タスクは，英語では前置詞誤りの訂正に相当する．英語の前置詞・冠詞誤りの訂正では，分類器を用いて適切な前置詞を選択するアプローチが多い (Gamon 2010; Han et al. 2010; Rozovskaya and Roth 2011)．これらは，誤りの種別を限定することにより，分類器による訂正を可能としている．一方，Mizumoto et al. (2011) は，日本語学習者の誤りの種別を限定せず，翻訳器を利用した誤り訂正を行った．この方法は，誤りを含む学習者作文を正しい文に変換することにより，あらゆる種類の誤りを訂正することを狙ったものである．本稿の訂正対象は助詞誤りであるが，今後の拡張性を考慮して，翻訳器と同様な機能を持つ識別的系列変換 (Imamura et al. 2011) をベースとした誤り訂正を行う．

翻訳の考え方を使った場合，モデル学習のために，誤りを含む学習者作文とそれを訂正した修正文のペア（以下，単にペア文とも呼ぶ）が大量に必要である．しかし，実際の学習者作文を大規模に収集し，さらに母語話者が修正するのはコストが高く難しい場合が多い．この問題に対し，本稿では以下の 2 つの提案を行う．

(1) 日本語平文コーパスの利用（言語モデル確率と二値素性の混在）
学習者作文・修正文ペアのうち，修正文側は正しい日本語であるため，既存の日本語平文コーパスなどから容易に入手可能である．そこで，比較的大規模な日本語平文コーパスを日本語修正文とみなして，変換器のモデルとして組み込む．組み込む際には，日本語平文コーパスは言語モデル確率の算出に利用し，学習者作文・日本語修正文ペアから獲得した二値素性と共に，識別モデルの枠組みで全体最適化を行う．学習者作文・修正文ペアに出現しないものであっても，言語モデル確率によって日本語の正しさが測られるため，誤り訂正の網羅性の向上が期待できる．

(2) 疑似誤り文によるペア文の拡張（とドメイン適応の利用）
学習者作文は容易に入手できないため，正しい文から誤りパターンに従って誤らせることにより，自動的に学習者作文を模した疑似誤り文を作成する．この疑似誤り文と元に

した日本語文をペアにして，訓練コーパスに追加する．ただし，自動作成した疑似誤り文は，実際の学習者作文の誤り分布を正確には反映していない．そのため，疑似誤りをソースドメイン，実誤りをターゲットドメインとみなして，ターゲットドメインへの適応を行う．疑似誤りの分布が実際の誤りと少々異なっていても，安定して精度向上ができると期待される．

以下，第 2 章では，我々が収集した日本語学習者作文の誤り傾向について述べる．第 3 章では，本稿のベースとなる誤り訂正法と，日本語平文コーパスの利用法について説明する．第 4 章では，疑似誤り文によるペア文の拡張法について説明し，第 5 章では実験で精度変換を確認する．第 6 章では関連研究を紹介し，第 7 章でまとめる．

2　日本語学習者の誤り傾向

まず，実際に外国人がどのような日本語書き誤りをしてしまうのか，日本語を学んでいる中国語母語話者を対象に誤り例を収集した．

被験者は日本語の学習歴があり，日本の技術系大学に在籍する，もしくは卒業した背景をもつ 37 名である．日本滞在歴は半年から 6 年程度である．各被験者に技術系文書（Linux マニュアル等 80 文）の英文と 24 個の図（のべ 104 課題）を提示し，キーボード入力による日本語作文を実施した（これを学習者作文と呼ぶ）．最終的には 2,770 文の学習者作文データを収集し，各作文を日本語母語話者が推敲した（以下，単に修正文と呼ぶ）．誤りを訂正する際には，文意を変更せず，文法的に正しい日本語とするための最小限の訂正を行うよう留意した[1]．言い換えると，この推敲で訂正された誤りは，訂正しないと正しい日本語にはならないものである．

2.1　誤りの分類と出現分布

誤り傾向の分析にあたり，まずは大分類として，文法誤り，語彙誤り，表記誤りの 3 種類を設定し，さらに小分類を設定した（表 1）．

収集した 2,770 文の分析を実施したところ，訂正が可能であったものは 2,171 文であった．訂正が出来なかったものは，全く誤りがない日本語文 559 文，および文として不完全な断片 40 文である．これ以降の分析は，訂正が可能であった 2,171 文に対して行った．

まず，誤り訂正の発生箇所は 4,916 箇所であり，1 文あたり平均 2.26 箇所であった．また各誤りの種別について，誤り大分類での出現分布をみると，文法誤りが 54%と最も多く，続いて語彙誤り 28%，表記誤りが 16%であった．これ以外は複数の誤りが混在する複合型誤りである．さらに小分類での出現分布をみると，最も多く発生していたのは助詞・助動詞誤り 33%，続い

[1] ただし，用語の選択誤りは訂正した．

表 1 誤りの分類と誤り例

大分類	小分類	誤り例
文法誤り	助詞・助動詞，活用，接続詞，指示詞，疑問詞，語順，態，時制	[助詞] 質問**を**対応する [活用] ブックを**開きて**ください [接続詞] 20 MB を超える**だから**アップロードします [指示詞] **その**以下のサイズに設定
語彙誤り	同音異義語，単語選択（類義語），母語の混用	[同音異義語] メモリ**内臓** [類義語] **快速**に処理します
表記誤り	カタカナ語，促音長音濁音，誤字脱字	[カタカナ語] アイコンを**クリーク**する [促音長音濁音] 質問が**あたら**お願いします [誤字脱字] 私**立ち**でやります

てカタカナ語誤り 11%，単語選択（類義語）の誤り 10%であった．

2.2 誤り傾向

今回の誤り傾向であるが，助詞誤りおよびカタカナ誤りは中国語母語話者に限らず広く外国人に共通して出現するものであると推測される．助詞は日本語特有の文法であり，多くの非日本語母語話者にとっては習得が難しいものである．そのため，中国語母語話者に限らず外国人の学習者作文の誤りに対する訂正対象を助詞とすることは，発生率から考えても効果的である．

助詞の種類によって誤り発生のしやすさは異なっているはずであり，全ての助詞が一律に誤りとはならない．今回の作文データにおける助詞誤りについて，さらに詳細に内訳を分析をしたところ，まず，誤りタイプとしては置換誤りが 74%，助詞の抜けが 17%，余分な助詞の出現が 9%であった．特に置換誤りの発生が高い．また余分な助詞の出現が 9%と非常に低く，訂正のために助詞の削除操作が必要となるケースは少ないことがわかる．個別の助詞誤り発生回数上位 10 件は表 2 のとおりである．

このうち，「は→が」への置換訂正については，1 文中に 2 回，「は／係助詞」が出現し，片方を「が／格助詞」に置換しなければならなかったものである（たとえば，「問題はあるときは...」）．「の」の助詞抜けとしては，「2 つファイル」のように，数量表現に後続する名詞の直前の「の」が欠けている誤りがよく見られた．また，余分な助詞「の」としては，「やったの人」「小さいの絵」など，連体修飾で使用された動詞や形容詞に後続して「の」が余分に存在している誤りが多い．

以上の分析から，本稿では，誤りの出現頻度の高い助詞誤りを訂正対象とした．また，助詞の置換，挿入，削除が現れていることから，原文（入力文）を置換，挿入，削除操作することにより，誤り訂正を行う．

表 2　頻出した助詞誤り

誤り	正解	訂正タイプ	頻度
は／係助詞	が／格助詞	置換	117
を／格助詞	が／格助詞	置換	87
	の／連体化	挿入	70
を／格助詞	に／格助詞	置換	69
を／格助詞	が／格助詞	置換	66
	を／格助詞	挿入	65
が／格助詞	は／係助詞	置換	65
の／連体化		削除	61
は／係助詞	を／格助詞	置換	54
	に／格助詞	挿入	49

3　識別的系列変換

本章では，ベースとなる識別的系列変換を用いた誤り訂正方式について述べる．本稿の誤り訂正は，学習者作文および修正文をあらかじめ形態素解析し，単語列から単語列へ変換することで行う．本方式は，基本的には識別モデルを用いた句に基づく統計翻訳器と同等であるが，挿入，削除操作への拡張と，言語モデル確率を扱う拡張を行っている．分類器を用いる誤り訂正方法と異なり，1 文中の複数の誤りを一度に訂正し，助詞以外の誤りにも拡張が可能な方式である．

3.1　基本方式

本稿では，音声認識結果を言語処理用単語列に変換する形態素変換器 (Imamura et al. 2011) をベースにし，以下の手順で入力文の誤りを訂正する．

- まず，入力単語列でフレーズテーブルを検索し，入力側にマッチするフレーズを得る．フレーズテーブルは，助詞誤りとその訂正候補を対にして格納したものである．これは誤り訂正タスクにおける Confusion Set (Rozovskaya and Roth 2010a) と同じもので，表 2 をテーブル化したものである[2]．フレーズテーブルと照合することにより，すべての訂正候補が得られる．また，無修正の場合を考慮し，入力単語を出力単語にコピーしたフレーズを作成し，両者をまとめてラティス構造にパックする（図 1）．これをフレーズラティスと呼ぶ．
- フレーズラティスから，条件付き確率場 (Conditional Random Fields; CRF) (Lafferty et al. 2001) に基づき，最尤フレーズ列を探索する．本稿の誤り訂正では語順の変更を行

[2] 表 2 はフレーズテーブルの一部である．5.1 節で述べるように，実際には ipadic-2.7.0 の最上位品詞が「助詞」であるすべての単語間の誤りを対象とした．

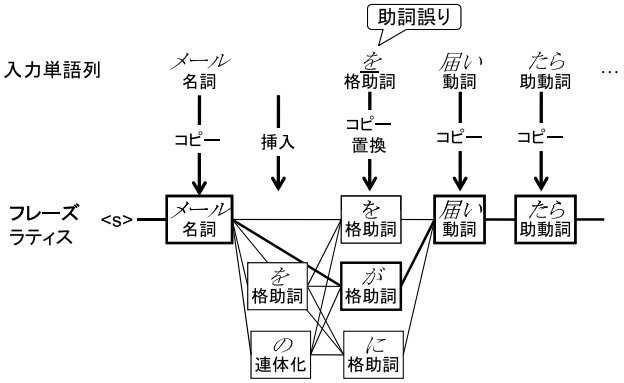

図 1　フレーズラティスの例（太線は正解系列を表す）

わないため，探索には Viterbi アルゴリズムを用いる．フレーズラティスには，非文法的系列（たとえば，図 1 では，格助詞「を」が連続する系列も候補として存在）も含まれるが，枝刈りなどは行わず，モデルに従い最尤探索を行う．

- 学習時には，学習者作文と修正文に対して，DP マッチによる単語アライメントを行い，正解のフレーズ列を作成する．この正解から，助詞誤りだけを取得してフレーズテーブルを作成するほか，正解を教師データとして CRF を学習する[3]．

3.2　挿入・削除操作

一般的に句に基づく翻訳器は置換操作のみで翻訳を行うが，本稿で実施する誤り訂正は，助詞の置換操作のほかに，挿入，削除操作も対象となる．挿入操作は，空単語からある単語への置換，削除操作は，ある単語から空単語への置換とみなせるため，両者も基本的には置換操作と同等に扱い，モデルの学習・適用を行う．

しかし，挿入操作は，全単語間に挿入される可能性があるため，ラティス構築時にサイズが爆発するなど，非常に計算コストの高い操作である．挿入箇所をある程度絞ることが望ましいため，本稿では，名詞直後に後続する助詞のみ，挿入を許可するという制約をかける．挿入は 1 箇所 1 単語のみとする．この制約により，一部訂正不可能な誤りも生じる（たとえば，格助詞「に」の直後に係助詞「は」を挿入し，「に」を「には」に訂正するのは不可能となる）．

なお，置換操作は，挿入操作と削除操作の連続でも表現できる．本稿では，挿入と削除操作が連続していた場合は，置換操作になるように正解データを作成し，モデルを学習する．誤り訂正時には，フレーズラティス内に置換操作の候補と，挿入と削除操作が連続する候補が混在

[3] 本稿では，CRF 学習のための最適化プログラムとして岡崎の libLBFGS を用い，実装した．
http://www.chokkan.org/software/liblbfgs/

するが，誤り訂正モデルに従い最尤探索すると，ほとんどすべての場合，置換操作が選ばれる．

3.3 素性

本手法では 2 種類の素性を用いる．一つは翻訳モデルに相当する入力と出力のフレーズ対応度を測るためのマッピング素性，もう一つは言語モデルに相当する出力単語列の日本語としてのもっともらしさを測るためのリンク素性である．マッピング素性とリンク素性の概要を図 2 に，素性テンプレートの一覧を表 3 に示す．

固有表現抽出など，識別モデルを用いるタスクでは，タグを付与すべき単語のほかに，その周辺単語を素性として用いる場合が多く，今回も同様な考え方をする．具体的には，当該フレー

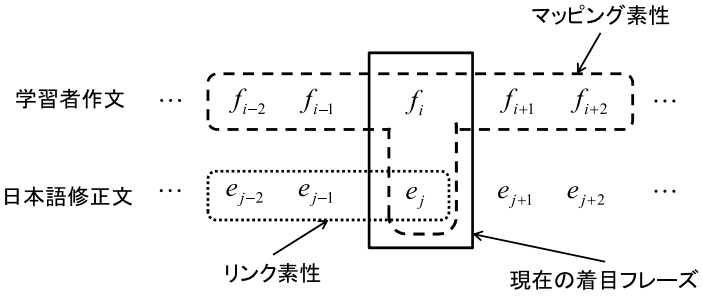

図 2　マッピング素性とリンク素性

表 3　素性テンプレート

種別	No.	関数値	内容	備考
マッピング	1	二値	$\delta(e_j, f_{i-2})$	
	2	二値	$\delta(e_j, f_{i-1})$	
	3	二値	$\delta(e_j, f_i)$	
	4	二値	$\delta(e_j, f_{i+1})$	
	5	二値	$\delta(e_j, f_{i+2})$	
	6	二値	$\delta(e_j, f_{i-2}, f_{i-1})$	
	7	二値	$\delta(e_j, f_{i-1}, f_i)$	
	8	二値	$\delta(e_j, f_i, f_{i+1})$	
	9	二値	$\delta(e_j, f_{i+1}, f_{i+2})$	
	10	二値	$\delta(e_j, f_{i-2}, f_{i-1}, f_i)$	
	11	二値	$\delta(e_j, f_{i-1}, f_i, f_{i+1})$	
	12	二値	$\delta(e_j, f_i, f_{i+1}, f_{i+2})$	
リンク	13	二値	$\delta(e_j)$	n-gram 二値素性
	14	二値	$\delta(e_j, e_{j-1})$	n-gram 二値素性
	15	二値	$\delta(e_j, e_{j-1}, e_{j-2})$	n-gram 二値素性
	16	実数値	$\log(P(e_j \mid e_{j-1}, e_{j-2}))$	言語モデル確率

表中の δ は，引数が完全一致したときに 1，それ以外は 0 を返す二値関数を表す．

ズの入力側前後 2 単語をウィンドウとして，1〜3-gram と当該フレーズの出力単語の対を，二値のマッピング素性として使用する．

リンク素性に関しては，次節で詳細に述べる．

3.4　日本語平文コーパスの利用とリンク素性への組み込み

誤り訂正タスクにおいては，「正しい日本語」を出力する必要があるため，リンク素性は重要であると考えられる．この「正しい日本語」は，既存の日本語平文コーパスから容易に入手可能である．そこで以下の 2 種類のリンク素性を併用し，識別学習を通じて全体最適化を行う．識別モデルを用いる本稿の方式は，相互に依存する素性を混在できるという特徴を利用している．

- n-gram 二値素性：出力単語の 1〜3-gram を二値素性として使用する．最適化用の訓練コーパス（学習者作文・修正文などのペア文）からしか獲得できない．個々の n-gram の素性重みは，マッピング素性を含む他の素性との兼ね合いを考慮しながら最適化されるため，きめ細かい最適化ができ，訓練コーパスにおける精度は高い．言い換えると，未知テキスト中に訓練コーパスと同じパターンの誤りが出現した場合，非常に高い精度で訂正ができる．
- 言語モデル確率：出力単語列の n-gram 確率（実際にはトライグラム確率）の対数値を実数素性として使用する．素性重みは 1 つしか付与されないが，言語モデルは日本語平文コーパスから学習できるため，訓練コーパスに限らず，大量の文から構築できる．訓練コーパスに出現した／しないにかかわらず，日本語としての適切さをスコアとして与えることができる．

識別学習における二値素性と実数素性の混在は，半教師あり学習における補助モデル (Suzuki et al. 2009; 鈴木，磯崎 2010) と同じ考え方であり，訓練コーパス上での精度を保ちながら，未知テキストに対して頑健な訂正が行えるという利点がある．

4　疑似誤り文を用いたペア文の拡張

第 3 章で述べた誤り訂正器には，学習のため，翻訳における対訳文に相当する学習者作文・修正文ペアが必要である．しかし，実際の誤り事例を大量に収集するのは困難であるため，自動生成した疑似誤り文を用いてペア文を拡張する．本章では，まず疑似誤り文生成方法について説明し，ドメイン適応を利用した疑似誤り文の適用方式について説明する．

4.1　疑似誤り生成

前述のとおり，学習者作文・日本語修正文ペアのうちの日本語修正文に関しては，日本語平文コーパスなどから文を適当に選択することにより，容易に入手できる．よって，収集した文

を，学習者作文のように誤らせることができれば，ペア文として扱うことができる．

本稿では，Rozovskaya and Roth (2010b) と同様の生成方法を取る．具体的には，フレーズテーブルには，すでに誤った助詞とその訂正候補が記録されているので，これを逆に適用し，訂正候補助詞が出現したら，正しい助詞を誤らせる．誤りはある確率で発生させるが，発生確率には，実誤りコーパス（学習者作文と日本語修正文ペア）上での正解助詞 e とその誤り助詞 f の相対頻度を使用する．すなわち，

$$P_{error}(f|e) = \frac{C(f,e)}{C(e)}, \tag{1}$$

ただし，$P_{error}(f|e)$ は誤り発生確率，$C(f,e)$ は，実誤りコーパス上での正解助詞 e とその誤り助詞 f の共起頻度，$C(e)$ は同コーパス上での正解助詞 e の出現頻度である．

このように生成した疑似誤り文を訓練コーパスに加えることにより，誤り訂正モデルを学習する．

4.2　素性空間拡張法によるドメイン適応

自動で作成した疑似誤り文の問題点は，実際の誤りの確率分布を反映している保証がない点である．より正確に実誤りに近づけるため，本稿ではドメイン適応の技術を用いる．すなわち疑似誤り文コーパスをソースドメイン，実際の学習者作文コーパスをターゲットドメインとみなし，ターゲットドメインに適応させた誤り訂正モデルを学習する．

本稿では，ドメイン適応法に Daume III (2007) の素性空間拡張法 (Feature Augmentation) を用いる．これは，素性空間を拡張することによりドメイン適応を行うもので，ソースドメインに関するモデルを事前分布と考えることに相当する．また，学習方法（学習器）を変更する必要がないという特徴がある．

素性空間拡張法を簡単に説明する．素性選択によって構築された素性は，共通，ソース，ターゲットの素性空間に拡張して配備される．この際，ソースドメインから作成された素性 (D_s) は共通およびソースに，ターゲットドメインから作成された素性 (D_t) は共通およびターゲットの素性空間に配備する．つまり，素性空間が 3 倍に拡張される（図 3）．

パラメータ推定は，上記素性空間上で通常どおり推定される．その結果，ソースドメイン，ターゲットドメインで共通に用いられる素性（つまり，ソース，ターゲットで矛盾しない素性）に関しては，共通空間の重みが大きくなり，両者で矛盾する素性に関しては，ソースまたはターゲット空間の素性が重くなる．どちらか片方にしか出現しない素性については，共通空間とドメイン依存空間の素性が重くなる．

図 3 には，素性空間拡張法の適用例も示した．ここでは，格助詞「が」を「を」に置換するか，無修正にするかという問題に単純化する．いま，ソースドメインデータ，ターゲットドメインデータから，以下の 3 種類の素性が得られたとする（表 3 の素性 No. 11 を想定）．

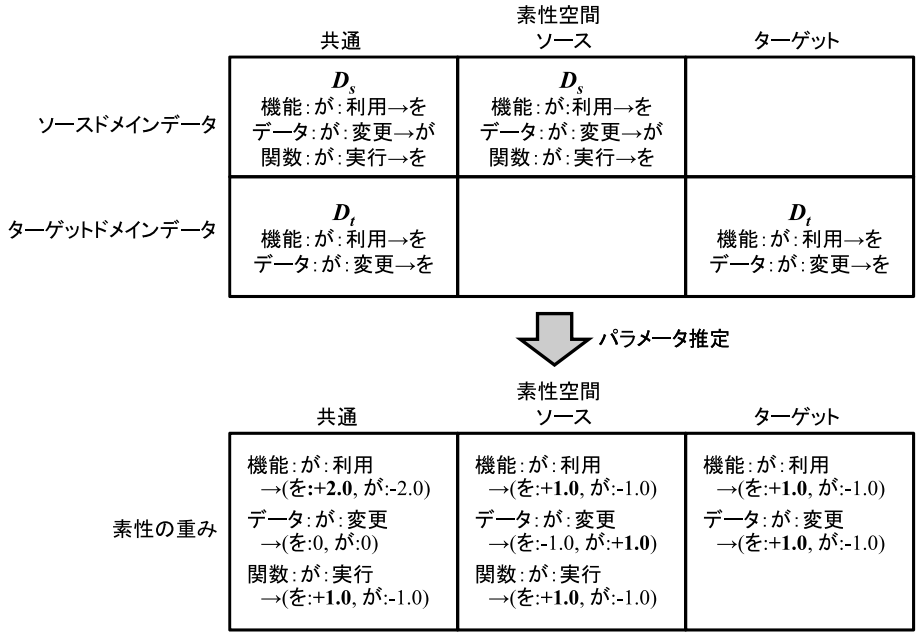

図 3　素性空間の拡張

- 「機能:が:利用」は，ソースドメイン，ターゲットドメイン双方に現れ，どちらも「を」に訂正している．
- 「データ:が:変更」は，ソース，ターゲット双方に現れているが，ソースドメインでは無修正，ターゲットドメインでは「を」に置換されている．
- 「関数:が:実行」は，ソースドメインのみに現れている．

この素性空間上でパラメータ推定を行うと，「機能:が:利用」は，ドメイン間で矛盾しないので，共通空間の重みが特に大きくなる．一方，「データ:が:変更」は，ソース・ターゲットで矛盾しているので，共通空間の重みが 0 になり，ソースまたはターゲット空間で，訂正先に依存した重みが重くなる．また，「関数:が:実行」は，共通空間とソース空間の重みが大きくなっている．

誤り訂正時には，共通とターゲット空間の素性のみを利用してデコードが行われる．ターゲットドメインに最適化されているため，実際の誤り出現分布に近くなる．また，ターゲットドメインの訓練データに現れない素性に関しても，ソースドメインデータから学習された共通空間の素性が利用できるため，ターゲットドメインのみを利用するときより，未知の入力に頑健になる．図 3 の例では，ソースドメインのみに出現した「関数:が:実行」も利用して訂正ができる．

5　誤り訂正実験

5.1　実験設定

訂正対象助詞：本稿で誤り訂正の対象とする助詞は，ipadic-2.7.0 の最上位品詞が助詞であるものすべてである．これには，格助詞，係助詞のほか，副助詞，接続助詞，終助詞，並立助詞なども含まれ，のべ 236 種類あるが，後述する学習者作文コーパスに出現しない，もしくは誤りがない助詞は訂正対象にならないため，実際の訂正対象助詞は 38 種類である[4]．

学習者作文コーパス（実誤りコーパス）：実験に使用したコーパスは，2 章で述べた 2,770 文（104 課題）である．ここから助詞誤りのみを残し，それ以外の部分は日本語修正文の単語を埋め込んだ文を作成，コーパスとした．つまり，実験に使用したペア文は，助詞誤りのみを含んだものである．ただし，助詞の表記が学習者作文，日本語修正文で一致している場合は誤りとはみなさず，日本語修正文の品詞を学習者作文にコピーして利用した．誤り総数は，助詞 13,534 個中 1,087 箇所 (8.0%) である．また，誤り助詞と訂正助詞を対にした異なり数は，132 種類（置換修正 95 種類，挿入 14 種類，削除 23 種類）である．なお，実験に使用したすべての文は，MeCab[5]（辞書は ipadic-2.7.0 を使用）によって形態素解析し，その表記と品詞を単語情報とした．

言語モデル：言語モデルは，Wikipedia のコンピュータ関連記事と，CentOS 5 の日本語マニュアルから，のべ 527,151 文を取得し，SRILM (Stolcke et al. 2011) でトライグラムを学習して使用した．バックオフ推定には，Modified Kneser-Ney ディスカウントと補間推定を併用し，未知ユニグラムを疑似単語 `<unk>`として残す設定で学習した．

疑似誤りコーパス：疑似誤り文は，言語モデル作成用コーパスから，ランダムに 10,000 文を取得して生成した．誤り発生確率は，実誤りコーパス上での相対頻度を倍率 1.0 として，倍率 0.0（つまり誤りなし）〜2.0 まで変化させて実験を行った．

評価法：評価は，コーパスを課題単位に分割し，5 分割交差検定で行った．評価基準は 2 種類使用した．

(1) 正解の単語列とシステム出力の単語列の表記を比較し，誤り訂正の再現率，適合率，F 値を算出した．

(2) 本タスクは，訂正すべき助詞数に比べ，訂正不要な助詞が圧倒的に多く，システムによって訂正不要な助詞を過剰に訂正してしまう懸念がある．そのため，訂正によって文の品質が向上した助詞数（訂正が必要な助詞をシステムが正しく訂正した数）と悪化した助詞数（訂正不要な助詞を過剰に訂正した数）の差を相対向上数として評価基準とした．こ

[4] Suzuki and Toutanova (2006) が対象とした格助詞 10 種+係助詞「は」と比べると，本稿では「へ」が対象外，「より」は単独では出現せず，連語「により」が訂正対象となっている．また，上記論文では，「に／は」のように，格助詞と「は」の連続も 1 語扱いで訂正対象としているが，本稿では連続した置換，挿入，削除操作を用いて訂正している．

[5] http://mecab.sourceforge.net/

の基準では，まったく修正を行わなかった場合に ±0 となる．

5.2 実験結果 1: 日本語平文コーパスの利用

まず，日本語平文コーパスを言語モデル確率として利用することの効果を測るため，以下の 3 手法について精度測定を行った．

- **提案手法**：リンク素性に n-gram 二値素性，言語モデル確率を併用した場合．
- **n-gram 二値素性のみ**：リンク素性に n-gram 二値素性のみを用い，言語モデル確率を使用しない場合．
- **言語モデル確率のみ**：リンク素性に言語モデル確率のみを用い，n-gram 二値素性を使用しない場合．

実験結果を表 4 に示す．表中の † は提案手法と n-gram 二値素性のみの間で有意差があったもの，§ は提案手法と言語モデル確率のみの間で有意差があったものを表す ($p < 0.05$) [6]．

まず適合率について，使用したリンク素性を比較すると，提案手法と n-gram 二値素性のみが同じ精度で，言語モデル確率のみの適合率が低めとなった．再現率は，提案方式が他の 2 つの方法に比べて大幅に向上 (9.9%，11.2%→ 18.9%) し，その結果，F 値も高い値を示した．n-gram 二値素性のみと言語モデル確率のみを比較すると，言語モデル確率のみの方が若干再現率が高い．その結果，F 値は提案手法（両者併用），言語モデル確率のみ，n-gram 二値素性のみの順で精度が高くなった．

しかし，相対向上数をみると，言語モデル確率のみは若干悪化しており（つまり過剰訂正が多い），再現率の向上が，誤り訂正の精度に直結していないことがわかる．これは，約 92%の助詞を無訂正にすべきという本タスクの特徴に由来するもので，安易な再現率向上は過剰訂正を引き起こすことを示している．

提案手法は，相対向上数でも他の 2 方式に勝っている．ただし，提案手法と n-gram 二値素性のみの間では有意差はなかった．これは，n-gram 二値素性は確実な誤りに集中して訂正する効果があるためで，相対向上数からみると有利に働いたためと考えられる．提案方式は，n-gram 二値素性，言語モデル確率の併用によって，適合率を保持したまま再現率を向上させており，誤

表 4 リンク素性を変えたときの誤り訂正結果

リンク素性	適合率	再現率	F 値	相対向上数
提案手法	**50.2%** (205/408)	**18.9%**†§ (205/1087)	**0.274**	**+28**§ (205 − 177)
n-gram 二値素性のみ	**50.2%** (108/215)	9.9% (108/1087)	0.166	+13 (108 − 95)
言語モデル確率のみ	46.4% (122/263)	11.2% (122/1087)	0.181	-6 (122 − 128)

[6] 適合率・再現率には比率の χ^2 検定を使用し，相対向上数には文単位の t 検定を使用した．

り訂正精度の向上に有効である．

5.3　実験結果 2: 疑似誤り文によるペア文の拡張

次に，疑似誤り文の導入効果を測定する．リンク素性を提案方法に限定し，疑似誤り文の使用方法のみを変えて実験を行う．

図 4 は，訓練に用いるコーパスと訓練法を以下の 4 通りに変えて，再現率／適合率カーブを測定した結果である．なお，図 4 は，誤り訂正器が出力するスコアが高い方から，ある再現率を達成するための訂正助詞を取得，適合率を算出したものである．

- **TRG:** 実誤りコーパスだけを用いて誤り訂正モデルを作成した場合（ベースライン）．
- **SRC:** 疑似誤りコーパスだけを用いて誤り訂正モデルを作成した場合．
- **ALL:** 実誤りコーパスに疑似誤りコーパスを単純追加してモデルを作成した場合．
- **AUG:** 提案方法．疑似誤りコーパスをソースドメイン，実誤りコーパスをターゲットドメインとして素性空間拡張法によるドメイン適応を行った場合．

TRG をベースラインと考えると，疑似誤り文のみ (SRC) では TRG の精度に達していない．そのため，疑似誤り文を追加した ALL でも適合率は再現率が高いところでようやく TRG と同等の適合率である．提案法である AUG は，再現率が高くなるに従い，TRG より高い適合率で誤りが訂正できている．再現率 18%では，TRG の適合率が 50.5%に対して，AUG の適合率は 55.4% となった（ただし，$p = 0.16$ で有意差はない）．なお，再現率 18%での SRC の適合率は 35.6%で，ランダムに訂正するのに比べると適合率は高い．

図 5 は，誤り発生確率毎の各方式の相対向上数をプロットしたものである．この実験では，誤り発生確率が低い方が全体的に精度がよく，誤り発生なし（倍率 0.0）から 0.6 までは ALL 方

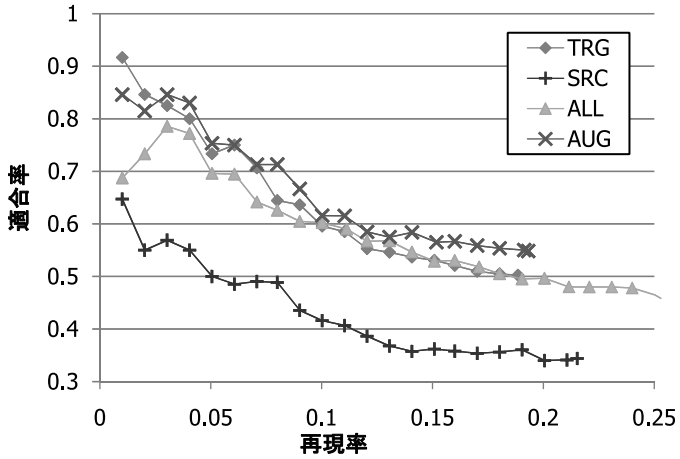

図 4　再現率／適合率カーブ（誤り発生確率は倍率 1.0 のとき）

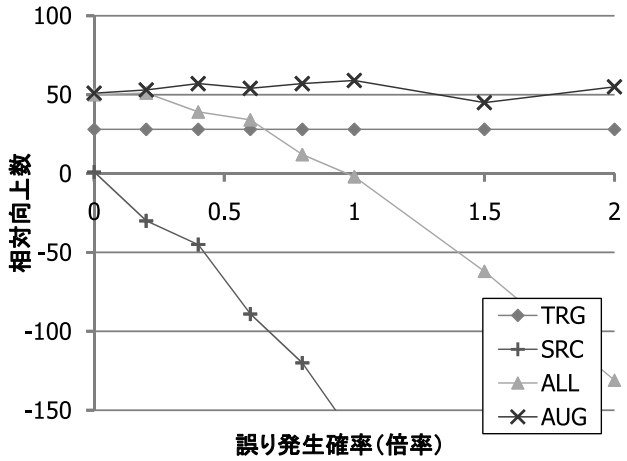

図 5 誤り発生確率（倍率）毎の相対向上数

式も TRG を上回っている．しかし，SRC は倍率を高くするに従って相対向上数が低下しており，誤り発生確率を適切に制御しないと，疑似誤り文が効果的に作用しない．一方 AUG は，誤り発生確率を変えても，安定した精度向上を果たした．誤り発生倍率が 1.0 のときの相対向上数は，TRG が+28 に対して AUG は+59 と，有意に向上しており，疑似誤り文を使用するときは，ドメイン適応を併用することが望ましい．

5.4 誤り訂正例

実験 2 において，誤り発生倍率 1.0 のとき，提案方式 (AUG) の適合率は 54.8% (210/383)，再現率は 19.3% (210/1087) であった．約 55%の適合率は，45%程度の修正箇所を再修正しないと正しい文にならないという意味で，実用上は決して高いとは言えない．助詞の用法には，意味的・文法的に明らかな誤用と，許容可能なものがあるため，人手評価を行った．

何らかの修正操作を出力したが，正解と異なった部分 173 箇所に関して，1 名の評価者によって主観評価した．なお，そのうち 151 箇所は，正解では無修正だった部分を過剰に修正したものである．評価観点は，システム修正を許容可能か（正解と比較して，意味的・文法的に異なっていないか）である．結果，173 箇所のうち 103 箇所は許容可能であった．つまり，許容可能という観点での適合率は，$(210 + 103)/383 = 81.7\%$ となった．

表 5 は，システムによる誤り訂正例である．置換，挿入，削除操作により誤り訂正が成功したもののほか，人手評価によって許容可能と判断されたものには，係助詞「は」と格助詞「が」の置換 (No. 4) や，複合名詞が正しい格助詞を補完して分割されたもの (No. 5) があった．許容不可として残ったものの中には，No. 7 のように慣用句を過剰訂正したもの，受動態をとらえられず，能動態の格助詞に置換したもの (No. 8)，Linux の free コマンドの内容を知らないと訂正

表 5　システムによる誤り訂正例（訂正部分周辺のみ）

タイプ	No.	入出力	文	備考
訂正成功	1	入力	これを押して、説明画面**が**表示します。	置換訂正
		出力	これを押して、説明画面**を**表示します。	
	2	入力	デザインの変更の要求＿あるので、やっていただきたい。	挿入訂正
		出力	デザインの変更の要求**が**あるので、やっていただきたい。	
	3	入力	新**の**システム稼動のため、移行が必要ですので、	削除訂正
		出力	新＿システム稼動のため、移行が必要ですので、	
許容可能	4	入力	準備**は**できているのでしょうか？	置換訂正
		出力	準備**が**できているのでしょうか？	
	5	入力	先ずはデザインのカード＿選択して下さい。	挿入訂正
		出力	先ずはデザインのカード**を**選択して下さい。	複合名詞分割
	6	入力	二つの数字の中＿、大きい数字を返します。	挿入訂正
		出力	二つの数字の中**の**、大きい数字を返します。	正解は「で」
許容不可	7	入力	申し訳ございません**が**、このメールを削除してください。	置換訂正
		出力	申し訳ございません**ので**、このメールを削除してください。	慣用句
	8	入力	「ファイルが破損しています」というメッセージ**が**表示されます。	置換訂正
		出力	「ファイルが破損しています」というメッセージ**を**表示されます。	受動態
	9	入力	私たちは試験をしたいです。あなた**は**いつ準備できますか？	置換訂正
		出力	私たちは試験をしたいです。あなた**が**いつ準備できますか？	呼応する表現
	10	入力	free は、カーネル**が**使ったバッファーを表示する。	置換訂正
		出力	free は、カーネル**を**使ったバッファーを表示する。	free の意味が必要

ができないもの (No. 10) があった．No. 9 は「は」と「が」の置換であるが，「私たち」と「あなた」が呼応する表現であるため，許容不可と判断された．本稿で用いた素性は訂正対象助詞の局所文脈のみであるため，大域的素性を導入しないと正しい訂正は困難なものもある．

6　関連研究

日本語学習者の助詞誤り検出･訂正は従来より研究されてきた. 近年では, Suzuki and Toutanova (2006) が，最大エントロピー法 (ME) による分類器を用いて，助詞（主に格助詞）が欠落した文からの復元を行っている．この入力文は形態素・構文解析済みであり，基本的に誤り箇所が既に分かっているとき，挿入操作だけで修正を行う．大木 他 (2011) は，形態素・構文解析済みの入力文（誤りを含む）に対して，周辺の形態素や係り先を素性として，SVM で助詞の誤用検出する方法を提案している．ここでは，助詞の欠落も対象としている．検出を行うのみで修正までは行わない．

英語の前置詞・冠詞誤り訂正では，Han et al. (2010) が，前置詞周辺単語や構文解析の主辞など

を素性とした ME 分類器を用いて，前置詞の誤り訂正を行った．Gamon (2010) は前置詞と冠詞誤りを対象に，ME 分類器による誤り検出，決定木による誤り訂正を行った．また，Rozovskaya and Roth (2010a) は平均化パーセプトロンに基づく分類器で前置詞の誤り訂正を行っている．これらの研究は，いずれも誤りの種類を助詞や前置詞・冠詞に限定することで，分類器による誤り訂正を可能としている．

一方，Mizumoto et al. (2011) は，誤りを助詞に限定せず，すべての誤りを対象とした自動訂正法を提案した．ここでは，対訳文に相当する学習者作文と日本人による修正文のペアを大量に SNS から収集し，句に基づく統計翻訳の仕組みを利用して訂正を行う．誤りを含む入力の形態素解析は行わず，文字単位で翻訳を行う．本稿で使用した系列変換は，基本的には統計翻訳と同等な手法である．そのため，誤りの種類を助詞に限定する必要がなく，他の誤りにも拡張できる．しかし，本稿の方式はあらかじめ学習者作文が単語に分割されていることを前提としている．誤りを含む文を形態素解析，構文解析した場合の精度は，一般的には日本語母語話者が記述した文の解析精度より落ちると考えられるため，単語分割法も併せて検討する必要がある (藤野 他 2012)．

母語話者の記述したテキスト（日本語修正文相当）のモデル化という観点で上記研究を俯瞰すると，Suzuki and Toutanova (2006), 大木 他 (2011), Han et al. (2010), Rozovskaya and Roth (2010a) は n-gram 二値素性として利用している．Gamon (2010), Mizumoto et al. (2011) は，n-gram 確率という形でモデル化している．本稿では，識別モデルの枠組みで両者を併用し，マッピング素性を含んで全体最適化を行うことにより，再現率を向上することができた．

学習者作文の利用という観点で俯瞰すると，いずれの研究も，学習者の誤り傾向をモデルとして組み込むことにより，母語話者の記述したテキストのみを用いて誤り訂正を行う場合に比べ，訂正精度が向上したと報告している (Han et al. 2010; Gamon 2010; Rozovskaya and Roth 2010a; 笠原 他 2012)．本稿の方式は，マッピング素性という形で学習者の誤り傾向をモデル化しており，従来研究の成果を取り込んでいる．

学習者作文を模した擬似誤り文に関しては，Rozovskaya and Roth (2010b) が提案を行っている．そこでは，学習者の実誤りと同じ分布を持つ擬似誤り文を追加することにより，精度が向上したと報告している．ただ，データ（論文では学習者の母語別）によって最適な擬似誤り生成方法が異なっており，擬似誤り生成を制御する必要がある．本稿では，擬似誤りと実誤りのずれをドメイン適応技術を用いて修正することで安定した精度向上ができた．

さまざまな種類の誤りの同時訂正は，Dahlmeier and Ng (2012) も行い，前置詞・冠詞誤りだけでなく，スペルミス，句読点，名詞の数の誤りも含めて訂正を行っている．誤りの種別ごとに分類器やルールを用いて訂正仮説を生成し，山登り的に書き換えを繰り返すことで 1 文中の複数の誤りを訂正する．彼らは，複数の仮説を保持することで，山登り時に局所解に陥る可能性を軽減しているが，本稿の方式はすべての仮説をフレーズラティスに持ち，Viterbi アルゴリ

ズムで最適な組み合わせを探索しているので，モデル上は最適な訂正結果であることが保証されている．

本タスクは，訂正すべき助詞に比べ訂正不要な助詞が圧倒的に多く，安易な再現率の向上は誤り訂正精度（相対向上数）の改善に直結しないと述べた．これはデータ不平衡問題 (Imbalanced Data Problem) と呼ばれ，機械学習を実タスクに適用するときの主要な問題の一つと認識されている（たとえば，サーベイ論文 (He and Garcia 2009) を参照）．この問題の解決方法には，少数派と多数派のデータを増減させることで平衡させる方法（サンプリング法）や，少数派の分類誤り（本タスクの場合，訂正誤り）と多数派の分類誤りに異なるコストを与えて学習する方法（ベイズリスク最小法）など，さまざまなものが提案されており，本タスクに適用できるか検討する必要がある．なお，本稿で提案した疑似誤り文は，実誤りの分布を変えないようにデータを増やすのが目的であるので，少数派データを増やす over-sampling 法とは異なる位置づけである．

7　おわりに

本稿では，中国語母語話者の日本語作文における，助詞誤り訂正法を提案した．誤り訂正タスクで難しいのは，誤りを含む実際の学習者作文とその修正文を入手することである．この問題に対して，本稿では，まず日本語平文コーパスを利用して，言語モデル確率とペア文から獲得した二値素性を識別モデルの枠組みで併用し，誤り訂正の再現率を向上させた．また，学習者作文を模した疑似誤り文を自動生成し，学習コーパスに追加した．ドメイン適応を併用することにより，誤り発生確率によらず，安定した精度向上ができることを示した．

本稿で用いた識別的系列変換は，助詞誤りに限定せず，すべての誤りを対象とすることができる．今後は，他の種類の誤り訂正にも拡張するのが課題である．

謝　辞

本研究の一部は，*the 50th Annual Meeting of the Association for Computational Linguistics* で発表したものである (Imamura et al. 2012)．本論文に関して，非常に有益なコメントをいただいた査読者の方々に感謝する．

参考文献

Dahlmeier, D. and Ng, H. T. (2012). “A Beam-Search Decoder for Grammatical Error Correction.” In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Lan-*

guage Processing and Computational Natural Language Learning (EMNLP-CoNLL 2012), pp. 568–578, Jeju Island, Korea.

Daume III, H. (2007). "Frustratingly Easy Domain Adaptation." In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL-2007)*, pp. 256–263, Prague, Czech Republic.

藤野拓也，水本智也，小町守，永田昌明，松本裕治 (2012). 日本語学習者の作文の誤り訂正に向けた単語分割. 言語処理学会第 18 回年次大会, pp. 26–29.

Gamon, M. (2010). "Using Mostly Native Data to Correct Errors in Learners' Writing." In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT-ACL 2010)*, pp. 163–171, Los Angeles, California.

Han, N.-R., Tetreault, J., Lee, S.-H., and Ha, J.-Y. (2010). "Using an Error-Annotated Learner Corpus to Develop an ESL/EFL Error Correction System." In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta.

He, H. and Garcia, E. A. (2009). "Learning from Imbalanced Data." *IEEE Transactions on Knowledge and Data Engineering*, **21** (9), pp. 1263–1284.

Imamura, K., Izumi, T., Sadamitsu, K., Saito, K., Kobashikawa, S., and Masataki, H. (2011). "Morpheme Conversion for Connecting Speech Recognizer and Language Analyzers in Unsegmented Languages." In *Proceedings of Interspeech 2011*, pp. 1405–1408, Florence, Italy.

Imamura, K., Saito, K., Sadamitsu, K., and Nishikawa, H. (2012). "Grammar Error Correction Using Pseudo-Error Sentences and Domain Adaptation." In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL-2012) Volume 2: Short Papers*, pp. 388–392, Jeju Island, Korea.

笠原誠司，藤野拓也，小町守，永田昌明，松本裕治 (2012). 日本語学習者の誤り傾向を反映した格助詞訂正. 言語処理学会第 18 回年次大会, pp. 14–17.

Lafferty, J., McCallum, A., and Pereira, F. (2001). "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data." In *Proceedings of the 18th International Conference on Machine Learning (ICML-2001)*, pp. 282–289, Williamstown, Massachusetts.

Mizumoto, T., Komachi, M., Nagata, M., and Matsumoto, Y. (2011). "Mining Revision Log of Language Learning SNS for Automated Japanese Error Correction of Second Language Learners." In *Proceedings of 5th International Joint Conference on Natural Language Processing (IJCNLP-2011)*, pp. 147–155, Chiang Mai, Thailand.

大木環美，大山浩美，北内啓，末永高志，松本裕治 (2011). 非日本語母国語話者の作成するシ

ステム開発文書を対象とした助詞の誤用判定．言語処理学会第 17 回年次大会発表論文集, pp. 1047–1050.

Rozovskaya, A. and Roth, D. (2010a). “Generating Confusion Sets for Context-Sensitive Error Correction.” In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP-2010)*, pp. 961–970, Cambridge, Massachusetts.

Rozovskaya, A. and Roth, D. (2010b). “Training Paradigms for Correcting Errors in Grammar and Usage.” In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT-ACL 2010)*, pp. 154–162, Los Angeles, California.

Rozovskaya, A. and Roth, D. (2011). “Algorithm Selection and Model Adaptation for ESL Correction Tasks.” In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Techologies (ACL-HLT 2011)*, pp. 924–933, Portland, Oregon.

Stolcke, A., Zheng, J., Wang, W., and Abrash, V. (2011). “SRILM at Sixteen: Update and Outlook.” In *Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU 2011)*, Waikoloa, Hawaii.

末永高志，松嶋敏泰 (2012). ベイズ決定理論にもとづく階層 N グラムを用いた最適予測法と日本語入力支援技術への応用．言語処理学会第 18 回年次大会, pp. 6–9.

Suzuki, H. and Toutanova, K. (2006). “Learning to Predict Case Markers in Japanese.” In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL 2006)*, pp. 1049–1056, Sydney, Australia.

鈴木潤，磯崎秀樹 (2010). 大規模ラベルなしデータを利用した係り受け解析の性能検証．言語処理学会第 16 回年次大会発表論文集, pp. 19–22.

Suzuki, J., Isozaki, H., Carreras, X., and Collins, M. (2009). “An Empirical Study of Semi-supervised Structured Conditional Models for Dependency Parsing.” In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP-2009)*, pp. 551–560, Singapore.

略歴

今村　賢治：1985 年千葉大学工学部電気工学科卒業．同年日本電信電話株式会社入社．1995 年～1998 年 NTT ソフトウェア株式会社．2000 年～2006 年 ATR 音声言語コミュニケーション研究所．2006 年より NTT サイバースペース研究所（現 NTT メディアインテリジェンス研究所）主任研究員．現在に至る．

主として自然言語処理の研究・開発に従事．博士（工学）．電子情報通信学会，情報処理学会，言語処理学会，ACL 各会員．

齋藤　邦子：1996 年東京大学理学部化学科卒業．1998 年同大学院修士課程修了．同年日本電信電話株式会社入社．現在，NTT メディアインテリジェンス研究所主任研究員．自然言語処理の研究・開発に従事．言語処理学会，情報処理学会各会員．

貞光　九月：2004 年筑波大学第三学群情報学類卒業．2009 年筑波大学大学院システム情報工学研究科コンピュータサイエンス専攻修了．同年，日本電信電話株式会社入社．以来，自然言語処理の研究開発に従事．現在，NTT メディアインテリジェンス研究所音声言語メディアプロジェクト研究員．情報処理学会，言語処理学会各会員．

西川　　仁：2006 年慶應義塾大学総合政策学部卒業．2008 年同大学大学院政策・メディア研究科修士課程修了．同年日本電信電話株式会社入社．現在 NTT メディアインテリジェンス研究所にて自然言語処理技術の研究開発に従事．奈良先端科学技術大学院大学博士後期課程在学中．言語処理学会，人工知能学会，情報処理学会各会員．

（2012 年 4 月 18 日　受付）
（2012 年 7 月 15 日　再受付）
（2012 年 7 月 26 日　採録）